HandGCAT: Occlusion-Robust 3D Hand Mesh Reconstruction from Monocular Images

Shuaibing Wang[1,2]   Shunli Wang[1,2]   Dingkang Yang[1,2]
Mingcheng Li[1,2]   Ziyun Qian[1,2]   Liuzhen Su[1,2]   Lihua Zhang[1,2,3,4*]
[1]Academy for Engineering and Technology, Fudan University   [2]Institute of Meta-Medical, IPASS
[3]Jilin Provincial Key Laboratory of Intelligence Science and Engineering, Changchun, China
[4]AI & Unmanned Systems Engineering Research Center of Jilin Province, China
21210860016@m.fudan.edu.cn, lihuazhang@fudan.edu.cn

Abstract—We propose a robust and accurate method for reconstructing 3D hand mesh from monocular images. This is a very challenging problem, as hands are often severely occluded by objects. Previous works often have disregarded 2D hand pose information, which contains hand prior knowledge that is strongly correlated with occluded regions. Thus, in this work, we propose a novel 3D hand mesh reconstruction network HandGCAT, that can fully exploit hand prior as compensation information to enhance occluded region features. Specifically, we designed the Knowledge-Guided Graph Convolution (KGC) module and the Cross-Attention Transformer (CAT) module. KGC extracts hand prior information from 2D hand pose by graph convolution. CAT fuses hand prior into occluded regions by considering their high correlation. Extensive experiments on popular datasets with challenging hand-object occlusions, such as HO3D v2, HO3D v3, and DexYCB demonstrate that our HandGCAT reaches state-of-the-art performance. The code is available at https://github.com/heartStrive/HandGCAT.

Index Terms—3D hand mesh reconstruction, hand-object occlusion, computer vision

Fig. 1. The main idea of the proposed HandGCAT is to exploit the hand prior knowledge to imagine occluded regions.

I. INTRODUCTION

3D hand mesh reconstruction plays an important role in many applications, such as virtual reality, human-computer interaction, sign language translation, and robotics [1], [2]. It is now an increasingly popular task in computer vision, benefiting from advancements in deep learning algorithms [3]–[13]. As we all know, the RGB camera is lower cost and computationally friendly, and it is highly desired to recover 3D hand mesh from a single RGB image [1], [14]. Self-occlusion in hand is an inherent problem caused by the diversity of hand poses and shapes. When interacting with objects, hands exhibit even greater occlusion from almost any point of view [15]. However, eliminating severe occlusion is challenging and remains unsolved.

Previous methods [16]–[18] have attempted to alleviate occlusion problem through data augmentation. However, these augmentation methods are limited in the wild due to significant differences between synthetic and real occlusions. More recently, [19], [20] exploited the attention mechanism to remove the interference of occlusions. Although the attention-based methods have shown noticeable results under the occlusions,
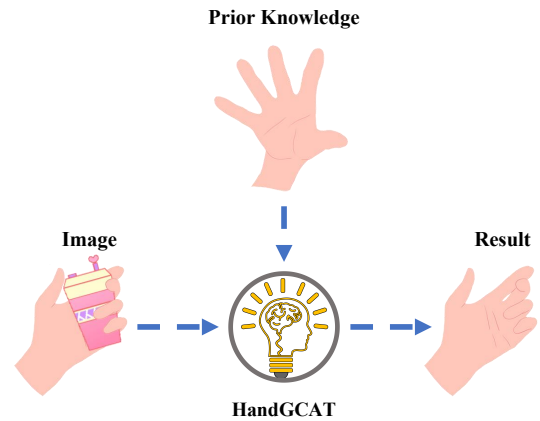
there are several limitations. First, these methods cannot imagine occluded regions. Second, they struggle to distinguish cluttered backgrounds that are unrelated to hands.

To resolve the above issues, let's think carefully about how humans solved this. Humans can easily imagine the appearance of the hand in occluded regions since they know exactly what the hand looks like. Inspired by this, we propose the HandGCAT, a novel framework for occlusion-robust 3D hand mesh reconstruction. As shown in Fig. 1, the main idea of the HandGCAT is to imagine occluded regions with prior knowledge. The main components of our network include a Knowledge-Guided Graph Convolution (KGC) module and a Cross-Attention Transformer (CAT) module. Guided by a predefined hand skeleton, KGC learns to hand prior knowledge from 2D hand pose based on a Graph Convolutional Network (GCN). CAT leverages hand prior knowledge to find relevant occluded regions and fuses prior information into occluded regions, which allows the network to imagine occluded regions as humans do. It has two advantages over existing methods. First, HandGCAT exploits 2D hand prior knowledge to compensate for the missing information in occluded regions. In addition, 2D hand pose can be estimated accurately from color images since many well-performing methods [21], [22] are
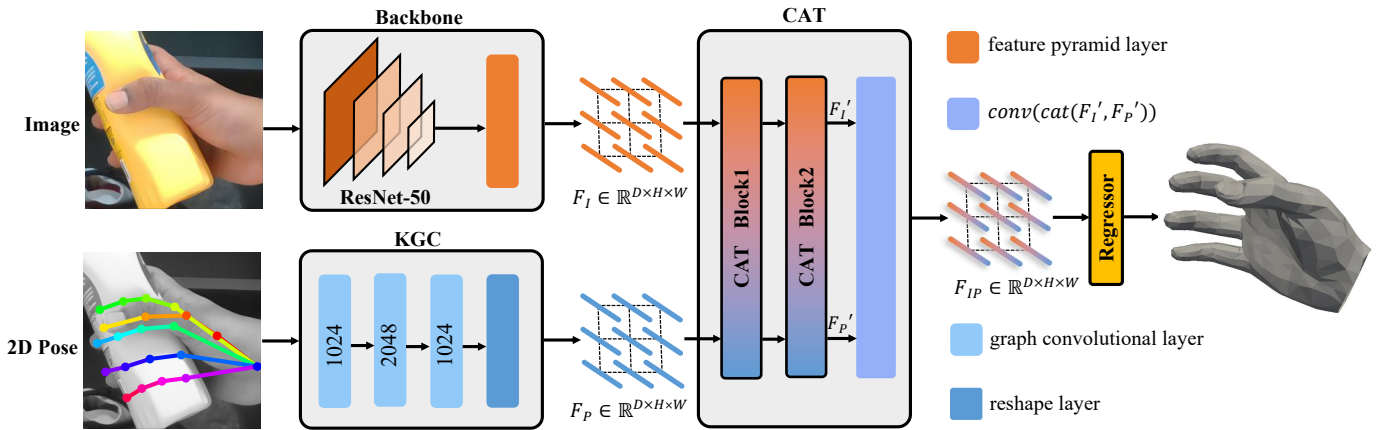
---

* Corresponding author

Fig. 2. Overview of the proposed HandGCAT for 3D hand mesh reconstruction, which consists of backbone, KGC, CAT, and regressor. Resnet-50 with FPN extracts image feature F_I. KGC captures hand prior knowledge F_P using GCNs from the 2D pose. CAT fuses F_P into F_I and thus imagines occluded regions. Finally, the regressor reconstructs the 3D hand mesh.

trained on real large-scale datasets. Second, HandGCAT fuses image features and hand prior knowledge for reconstruction, which can effectively alleviate the overfitting related to image appearance. For evaluation, we report the model's performance on three challenging benchmarks containing severe occlusions: HO3D v2 [23], HO3D v3 [24], and DexYCB [25]. Without whistles and bells, the HandGCAT can outperform the results of state-of-the-art methods.

Our main contributions are summarized as follows: (1) We propose a novel framework, HandGCAT, that recovers 3D hand mesh from a single RGB image. It is robust to severe hand occlusions and free from overfitting to image appearance. (2) To obtain hand prior knowledge and fuse it into occluded regions, We propose KGC and CAT modules. Based on a predefined hand skeleton, the KGC leverages the GCN to model the interactions of 2D hand joints. The CAT adopts hand prior to compensate for hand occluded regions and improves the accuracy of reconstructing hand mesh under occlusions. (3) Extensive experiments show our framework significantly outperforms state-of-the-art 3D hand mesh reconstruction methods on hand-object interaction datasets that contain severe hand occlusions.

II. RELATED WORKS

A. 3D Hand Mesh Reconstruction

Most of the 3D hand mesh reconstruction methods use a parametric model such as MANO [26] to represent the hand pose and shape. [17], [18] utilized data augmentation during the training time to handle the occlusion. [19], [20] applied the attention mechanism to improve the robustness under occlusions. In this work, we propose a knowledge-guided occlusion robustness strategy that achieves state-of-the-art reconstruction results compared to previous methods.

B. Graph Convolutional Networks

Compared with traditional Convolutional Neural Networks (CNN), Graph Convolutional Networks (GCN) can perform convolutional operations on graph data. As the skeleton can be represented in the form of a graph, GCNs are naturally widely used in 3D mesh reconstruction. [27], [28] adopted GCN to model the topological relationships between mesh vertices and achieved state-of-the-art performance. In this paper, we propose a knowledge-guided graph convolution module based on GCN to learn the implicit topology of the hand.

C. Transformers

Recently, many Transformer-based methods have been proposed for reconstructing 3D hand mesh from RGB images. [29] proposed Transformer-based networks to model global vertex-to-vertex interactions. [30] adopted a Transformer along with GCN to jointly model vertex-to-vertex and vertex-to-joint interactions. However, how to effectively exploit the hand's implicit topology to compensate for occluded regions has rarely been explored. In contrast, we propose a cross-attention Transformer that fuses hand features and implicit topology representations.

III. HANDGCAT

A. Overview

The network architecture is shown in Fig. 2. Given a hand image $I \in \mathbb{R}^{256 \times 256 \times 3}$, Resnet-50 with FPN is used as a feature extractor to obtain a feature map $F_I \in \mathbb{R}^{256 \times 32 \times 32}$. Then, the KGC module takes the 2D pose $P_{2D} \in \mathbb{R}^{21 \times 2}$ as input to perform joint-to-joint modeling, resulting in hand prior feature $F_P \in \mathbb{R}^{21 \times 32 \times 32}$ with rich topologic information. Note that in training time, we adopt the ground truth 2D pose, and in testing time, the 2D pose is estimated from a pretrained keypoint detector [20]. However, the estimated 2D pose often contains errors. To enhance robustness to errors, we added randomly generated noise to the input ground truth 2D pose. In the occlusion-aware stage, the CAT module fuses the corresponding hand prior into occluded regions, thus imagining the hand occluded regions, and finally outputs fused feature $F_{IP} \in \mathbb{R}^{256 \times 32 \times 32}$. In the reconstruction stage, the
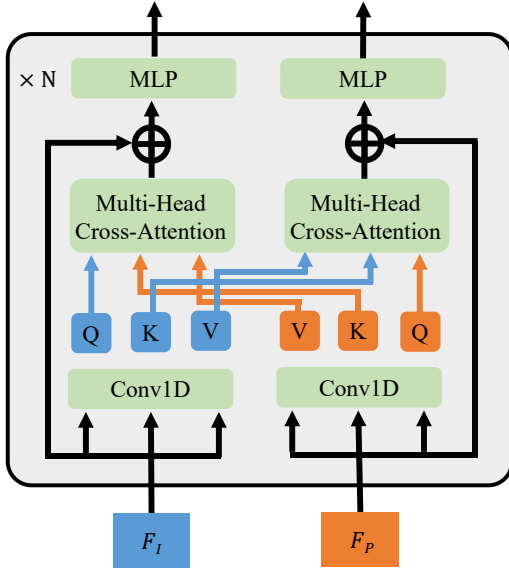
Fig. 3. Cross-Attention Transformer Block.

regressor takes the fused features F_{IP} as input to predict the 3D vertex coordinates.

B. Knowledge-Guided Graph Convolution Module

Humans have strong knowledge extraction and common-sense reasoning abilities. Therefore humans can easily imagine hand occluded regions. Inspired by this, we designed a Knowledge-Guided Graph Convolution module, which learns hand prior from 2D pose as complementary information for occluded regions.

Graph construction. We construct an undirected graph $\mathcal{G} = (\mathcal{V}, W)$ of P_{2D}, where $\mathcal{V} = \{p_i\}_{i=1}^J$ denotes a set of $J = 21$ hand joints, $W = (w_{i,j})_{J \times J}$ is the adjacency matrix. W defines the edge connections between joints based on the hand skeleton [26], where $w_{i,j} = 1$ if joint i and joint j are connected, and $w_{i,j} = 0$ otherwise. The normalized graph Laplacian is computed as $L_P = I_J - D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$, where I_J is the identity matrix, D is the degree matrix, and $D_{ii} = \sum_j W_{ij}$. The scaled Laplacian [28] is computed as $\tilde{L}_P = 2L_P/\lambda_{max} - I_J$.

Graph convolution on pose. The KGC module performs graph convolution operations on the 2D pose P_{2D} based on Chebysev spectral graph convolution [31], modeling joint and joint interactions to obtain hand prior information F_P. The formulation for performing the graph convolution is as follows.

$$F_{\text{out}} = \sum_{k=0}^{K-1} T_k\left(\tilde{L}_{\text{P}}\right) F_{\text{in}} \Theta_k \qquad (1)$$

where T_k is the Chebyshev polynomial of order k. We set K=1 in the KGC module, which means 1-hop neighbor nodes from each node are affected [28]. $F_{in} \in \mathbb{R}^{J \times f_{in}}$ and $F_{out} \in \mathbb{R}^{J \times f_{out}}$ are the input and output feature maps with dimensions f_{in} and f_{out}, respectively. $\Theta_k \in \mathbb{R}^{f_{in} \times f_{out}}$ is a

vector of Chebyshev coefficients.

C. Cross-Attention Transformer Module

In contrast to the conventional Transformer [38], we designed a Cross-Attention Transformer (CAT) module to imagine hand occluded regions with prior knowledge. As shown in Fig. 3, CAT utilizes multiple stacked CAT blocks to implicitly model the correlation between hand prior F_P and image feature F_I with occlusions, potentially transferring F_P to F_I.

We extract $Q_x, K_x, V_x, x \in \{I, P\}$ from F_I and F_P by two 1×1 convolution layers. Following the position encoding (PE) in Transformer [38], position information is added to Q_x and K_x:

$$Q_x = Q_x + PE(Q_x),$$
$$K_x = K_x + PE(K_x), \qquad (2)$$

Then, we use a multi-head cross-attention mechanism to perform the message passing between F_I and F_P. The bidirectional interaction of Q and K, V between F_I and F_P allows CAT to efficiently mine strong correlations between F_I and F_P. This process is denoted as follows:

$$F_{I \to P} = Q_P + \text{softmax}\left(\frac{Q_P K_I^T}{\sqrt{d}}\right) V_I,$$
$$F_{P \to I} = Q_I + \text{softmax}\left(\frac{Q_I K_P^T}{\sqrt{d}}\right) V_P, \qquad (3)$$

where $I \to P$ denotes pass F_I to F_P, $P \to I$ otherwise, d is the scaling factor. Afterward, we enhance the representation of the features using a multi-layer MLP as follows:

$$F_I' = MLP\left(F_I + F_{P \to I}\right),$$
$$F_P' = MLP\left(F_P + F_{I \to P}\right), \qquad (4)$$

F_I' and F_P' can be taken as input for the next CAT block. After deep stacking multiple CAT blocks, we finally utilize a 1×1 convolutional layer to obtain the fused features F_{IP}.

$$F_{IP} = Conv1D(concat(F_I', F_P')) \qquad (5)$$

D. Regressor and Loss Function

We follow the regressor from previous work [35]. The regressor consists of a single hourglass network [39], a parametric regression network, and a MANO layer. The hourglass network takes the fused features F_{IP} as input and outputs heatmaps $\mathbf{H} \in \mathbb{R}^{256 \times 32 \times 32}$ for each joint.

The parametric regression network concatenates the fused features F_{IP} with heatmaps as input and outputs MANO pose $\theta \in \mathbb{R}^{48 \times 3}$ and MANO shape $\beta \in \mathbb{R}^{10}$ parameters. The MANO layer produces 3D joints $J_{3D} \in \mathbb{R}^{21 \times 3}$ and 3D Mesh $V \in \mathbb{R}^{778 \times 3}$ based on MANO parameters.

To train our HandGCAT, we compute a loss function, which is defined as the L_2 distance between prediction $(\hat{\mathbf{H}}, \hat{\theta}, \hat{\beta}, \hat{J_{3D}}, \hat{V})$ and ground truth $(\mathbf{H}, \theta, \beta, J_{3D}, V)$.

TABLE I
COMPARISON WITH STATE-OF-THE-ART METHODS ON HO3D V2.

Method	PA-MPJPE ↓	PA-MPJPE AUC ↑	PA-MPVPE ↓	PA-MPVPE AUC ↑	F@5 ↑	F@15 ↑
I2L-MeshNet [32] (CVPR'20)	11.2	0.775	13.9	0.722	0.409	0.932
Hasson *et al.* [33] (CVPR'20)	11.0	0.780	11.2	0.777	0.464	0.939
Hampali *et al.* [34] (CVPR'20)	10.7	0.788	10.6	0.790	0.506	0.942
METRO [29] (CVPR'21)	10.4	0.792	11.1	0.779	0.484	0.946
Liu *et al.* [35] (CVPR'21)	9.9	0.803	9.5	0.810	0.528	0.956
I2UV-HandNet [2] (ICCV'21)	9.9	0.804	10.1	0.799	0.500	0.943
ArtiBoost [36] (CVPR'22)	11.4	0.773	10.9	0.782	0.488	0.944
Keypoint Trans. [14] (CVPR'22)	10.8	0.786	-	-	-	-
MobRecon [37] (CVPR'22)	9.2	-	9.4	-	0.538	0.957
HandOccNet [20] (CVPR'22)	9.1	0.819	8.8	0.819	0.564	**0.963**
HandGCAT (Ours)	**8.7**	**0.826**	**8.7**	**0.827**	**0.584**	**0.963**

TABLE II
COMPARISON WITH STATE-OF-THE-ART METHODS ON HO3D V3.

Method	PA-MPJPE ↓	PA-MPJPE AUC ↑	PA-MPVPE ↓	PA-MPVPE AUC ↑	F@5 ↑	F@15 ↑
ArtiBoost [36] (CVPR'22)	10.8	0.785	10.4	0.792	0.507	0.946
Keypoint Trans. [14] (CVPR'22)	10.9	0.785	-	-	-	-
HandOccNet [20] (CVPR'22)	10.7	0.786	10.4	0.791	0.479	0.935
HandGCAT (Ours)	**9.3**	**0.814**	**9.1**	**0.818**	**0.552**	**0.956**

IV. EXPERIMENTS

A. Implementation Details

All implementations were done with PyTorch. Adam optimizer is adopted to train the network with a mini-batch size of 32. All models in our experiments are trained for 70 epochs. The initial learning rate is 10^{-4}, which is scaled by a factor of 0.7 at every 10th epoch. All other details will be available in our codes.

B. Datasets and Evaluation Metrics

HO3D v2. HO3D v2 [23] is a hand-object interaction dataset with severe occlusions, which contains 66,034 training samples and 11,524 testing samples. For the training set, each sample provides a single-view RGB image, MANO-based hand joints and mesh, and camera parameters. For the testing set, each sample has an RGB image with the annotation of the detected bounding box. The results of the testing set can be evaluated through an online submission system.
HO3D v3. Compared to HO3D v2, HO3D v3 [24] has a larger scale, which provides 83,325 training samples and 20,137 testing samples. HO3D v3 provides more accurate annotations of hand-object poses, resulting in more contact between the hand and the object.
DexYCB. DexYCB [25] is a large-scale dataset of object grasping, containing 582K image frames of grasping on 20 YCB objects. We make use of the official "S0" split to evaluate the right-hand pose.
We report evaluation results using the following metrics.
MPJPE/MPVPE measures the average Euclidean distance (in mm) between the predicted joint/vertex coordinates and the ground truth coordinates.
PA-MPJPE/MPVPE is obtained by modifying MPJPE/MPVPE using Procrustes analysis.
AUC means the Area Under Curve (AUC) of the Percentage of Correct Keypoints (PCK) and Vertices (PCV) at different error thresholds.
F-Score is the harmonic mean between the predicted mesh and ground truth mesh at a specific distance threshold follow [35]. F@5/F@15 corresponds to the threshold of 5mm/15mm, respectively.

C. Comparisons with the State-of-the-art Methods

We compared the proposed method with several state-of-the-art methods to reconstruct hand mesh on the HO3D v2 and HO3D v3 datasets. In Tab. I&II, we can see that the proposed method outperforms the previous state-of-the-art for both datasets. To further validate the robustness of our method to occlusions, we compare it with state-of-the-art methods on a larger dataset DexYCB in Tab. III. Our method also outperforms all previous methods. In summary, the experimental results on three datasets all demonstrate the generalization capability of the proposed method to the situation where the hand is severely occluded.

D. Ablation Study

To prove the effectiveness of the HandGCAT method, we conduct further analysis on several essential components. The experimental results on the KGC module and CAT module are reported in Tab. IV&V, respectively.

TABLE III
COMPARISON WITH SOTA ON DEXYCB DATASET.

Method	MPJPE ↓	PA-MPJPE ↓
Spurr *et al.* [40] (ECCV'20)	17.34	6.83
METRO [29] (CVPR'21)	15.24	6.99
Liu *et al.* [35] (CVPR'21)	15.28	6.58
HandOccNet [20] (CVPR'22)	14.04	5.80
HandGCAT (Ours)	**13.76**	**5.60**

TABLE IV
COMPARISON OF MODELS WITH VARIOUS KGC ARCHITECTURES ON
HO3D v2.

KGC architectures	PA-MPJPE ↓	PA-MPVPE ↓	F@5 ↑	F@15 ↓
MLP	9.3	9.3	0.547	0.959
1-layer GCN	9.2	9.2	0.546	0.961
2-layer GCNs	9.0	8.9	0.570	0.961
3-layer GCNs	8.9	8.8	0.573	**0.963**
4-layer GCNs	**8.7**	**8.7**	**0.584**	**0.963**
5-layer GCNs	8.9	8.8	0.579	0.962

TABLE V
COMPARISON OF MODELS WITH VARIOUS CAT ARCHITECTURES ON
HO3D v2.

CAT architectures	PA-MPJPE ↓	PA-MPVPE ↓	F@5 ↑	F@15 ↑
Two Transformers	9.0	9.0	0.563	0.962
Single CAT block	8.9	8.8	0.574	0.962
Two CAT blocks	**8.7**	**8.7**	**0.584**	**0.963**
Three CAT blocks	8.8	**8.7**	0.583	**0.963**

Effectiveness of the KGC module. Tab. IV shows that our KGC module with 4-layer GCNs achieves the best results in all metrics, where MLP represents a 3-layer perceptron. By observing the Tab. IV, we can see that: (1) When using MLP instead of the KGC module, the error of the reconstruction is dramatically raised. This is because MLP struggles in learning the hand skeleton spatial structure. It demonstrated that the KGC module performs essential roles in learning 2D prior knowledge. (2) The KGC module achieves the best results with 4-layer GCN, and degradation occurs with 5-layer GCN. This is due to the KGC module with shallow GCN layers (*e.g.*, 1, 2, and 3 layers) being unable to model long-range joint interactions. However, the over-smooth problem of deeper GCN layers (*e.g.*, 5 layers) leads to performance degradation.

Effectiveness of the CAT module. Tab. V shows that our CAT module with two stacked CAT blocks achieves the best results in all metrics, where two transformers represent two layers of the traditional transformer. From Tab. V, we list the following observations: (1) With the transformer structure instead of the CAT module, which achieves slightly worse results. This suggests that the traditional transformer fails to capture the high correlation between prior knowledge and image features. (2) Two stacked CAT blocks are more effective than a single CAT block and three CAT blocks. This

Fig. 4. Qualitative comparison of the proposed HandGCAT and state-of-the-art method [20] on HO3D v2.

is because a single CAT block cannot adequately enhance the image feature, and an additional CAT block has a marginal effect on compensating for information in occluded regions.

E. Qualitative Analysis

Qualitative results on HO3D v2 are shown in Fig. 4. It can be seen that our HandGCAT produces the 3D mesh much better than the state-of-the-art method HandOccNet [20], even under severe occlusions. In the first example, the results generated by HandOccNet [20] are completely off from the angle of the original RGB image. In contrast, HandGCAT predicted relatively accurate results. With the help of hand prior knowledge, HandGCAT successfully imagined occluded regions, thus improving the performance of reconstruction.

V. CONCLUSION

In order to mitigate the challenges posed by hand occlusions, we propose a novel hand reconstruction method named HandGCAT, which is robust to occlusions. First, we propose a knowledge-guided graph convolution module to learn the hand prior. Second, we design a cross-attention Transformer to fuse the hand prior into occluded regions. Extensive experimental results show that our method achieves state-of-the-art performance on 3D hand mesh benchmarks that contain severe occlusions.

VI. ACKNOWLEDGMENT

REFERENCES

[1] Baowen Zhang, Yangang Wang, Xiaoming Deng, Yinda Zhang, Ping Tan, Cuixia Ma, and Hongan Wang, "Interacting two-hand 3d pose and shape reconstruction from single color image," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021.

[2] Ping Chen, Yujin Chen, Dong Yang, Fangyin Wu, Qin Li, Qingpei Xia, and Yong Tan, "I2uv-handnet: Image-to-uv prediction network for accurate and high-fidelity 3d hand mesh modeling," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021.

[3] Dingkang Yang, Shuai Huang, Haopeng Kuang, Yangtao Du, and Lihua Zhang, "Disentangled representation learning for multimodal emotion recognition," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, p. 1642–1651.

[4] Dingkang Yang, Haopeng Kuang, Shuai Huang, and Lihua Zhang, "Learning modality-specific and -agnostic representations for asynchronous multimodal language sequences," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, p. 1708–1717.

[5] Dingkang Yang, Shuai Huang, Shunli Wang, Yang Liu, Peng Zhai, Liuzhen Su, Mingcheng Li, and Lihua Zhang, "Emotion recognition for multiple context awareness," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022, vol. 13697, pp. 144–162.

[6] Dingkang Yang, Yang Liu, Can Huang, Mingcheng Li, Xiao Zhao, Yuzheng Wang, Kun Yang, Yan Wang, Peng Zhai, and Lihua Zhang, "Target and source modality co-reinforcement for emotion understanding from asynchronous multimodal sequences," *Knowledge-Based Systems*, vol. 265, pp. 110370, 2023.

[7] Haopeng Kuang, Dingkang Yang, Shunli Wang, Xiaoying Wang, and Lihua Zhang, "Towards simultaneous segmentation of liver tumors and intrahepatic vessels via cross-attention mechanism," *arXiv preprint arXiv:2302.09785*, 2023.

[8] Yuzheng Wang, Zhaoyu Chen, Dingkang Yang, Pinxue Guo, Kuaixun Jiang, Wenqiang Zhang, and Lizhe Qi, "Model robustness meets data privacy: Adversarial robustness distillation without original data," *arXiv preprint arXiv:2302.11611*, 2023.

[9] Dingkang Yang, Shuai Huang, Yang Liu, and Lihua Zhang, "Contextual and cross-modal interaction for multi-modal speech emotion recognition," *IEEE Signal Processing Letters*, vol. 29, pp. 2093–2097, 2022.

[10] Dingkang Yang, Zhaoyu Chen, Yuzheng Wang, Shunli Wang, Mingcheng Li, Siao Liu, Xiao Zhao, Shuai Huang, Zhiyan Dong, Peng Zhai, and Lihua Zhang, "Context de-confounded emotion recognition," 2023.

[11] Shunli Wang, Dingkang Yang, Peng Zhai, Chixiao Chen, and Lihua Zhang, "Tsa-net: Tube self-attention network for action quality assessment," in *Proceedings of the 29th ACM International Conference on Multimedia (ACM MM)*, 2021, pp. 4902–4910.

[12] Shunli Wang, Shuaibing Wang, Bo Jiao, Dingkang Yang, Liuzhen Su, Peng Zhai, Chixiao Chen, and Lihua Zhang, "Ca-spacenet: Counter-factual analysis for 6d pose estimation in space," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2022, pp. 10627–10634.

[13] Shunli Wang, Dingkang Yang, Peng Zhai, Qing Yu, Tao Suo, Zhan Sun, Ka Li, and Lihua Zhang, "A survey of video-based action quality assessment," in *2021 International Conference on Networking Systems of AI (INSAI)*, 2021, pp. 1–9.

[14] Shreyas Hampali, Sayan Deb Sarkar, Mahdi Rad, and Vincent Lepetit, "Keypoint transformer: Solving joint identification in challenging hands and object interactions for accurate 3d pose estimation," in *CVPR*, 2022.

[15] Tze Ho Elden Tse, Kwang In Kim, Ales Leonardis, and Hyung Jin Chang, "Collaborative learning for hand and object reconstruction with attention-guided graph convolution," in *CVPR*, 2022.

[16] István Sárándi, Timm Linder, Kai O Arras, and Bastian Leibe, "How robust is 3d human pose estimation to occlusion?," *arXiv:1808.09316*, 2018.

[17] Yu Cheng, Bo Yang, Bo Wang, and Robby T Tan, "3d human pose estimation using spatio-temporal networks with explicit occlusion training," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

[18] Lipeng Ke, Ming-Ching Chang, Honggang Qi, and Siwei Lyu, "Multi-scale structure-aware network for human pose estimation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[19] Lu Zhou, Yingying Chen, Yunze Gao, Jinqiao Wang, and Hanqing Lu, "Occlusion-aware siamese network for human pose estimation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[20] JoonKyu Park, Yeonguk Oh, Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee, "Handoccnet: Occlusion-robust 3d hand mesh estimation network," in *CVPR*, 2022.

[21] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu, "Rmpe: Regional multi-person pose estimation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2017.

[22] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh, "openrealtime multi-person 2d pose estimation using part affinity fields," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017.

[23] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit, "Honnotate: A method for 3d annotation of hand and object poses," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[24] Shreyas Hampali, Sayan Deb Sarkar, and Vincent Lepetit, "Ho-3d_v3: Improving the accuracy of hand-object annotations of the ho-3d dataset," *arXiv:2107.00887*, 2021.

[25] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S. Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, Jan Kautz, and Dieter Fox, "DexYCB: A benchmark for capturing hand grasping of objects," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[26] Javier Romero, Dimitrios Tzionas, and Michael J Black, "Embodied hands: Modeling and capturing hands and bodies together," *arXiv:2201.02610*, 2022.

[27] Liuhao Ge, Zhou Ren, Yuncheng Li, Zehao Xue, Yingying Wang, Jianfei Cai, and Junsong Yuan, "3d hand shape and pose estimation from a single rgb image," in *CVPR*, 2019.

[28] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee, "Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2020.

[29] Kevin Lin, Lijuan Wang, and Zicheng Liu, "End-to-end human pose and mesh reconstruction with transformers," in *CVPR*, 2021.

[30] Kevin Lin, Lijuan Wang, and Zicheng Liu, "Mesh graphormer," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021.

[31] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst, "Convolutional neural networks on graphs with fast localized spectral filtering," *Advances in neural information processing systems*, 2016.

[32] Gyeongsik Moon and Kyoung Mu Lee, "I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2020.

[33] Yana Hasson, Bugra Tekin, Federica Bogo, Ivan Laptev, Marc Pollefeys, and Cordelia Schmid, "Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction," in *CVPR*, 2020.

[34] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit, "Honnotate: A method for 3d annotation of hand and object poses," in *CVPR*, 2020.

[35] Shaowei Liu, Hanwen Jiang, Jiarui Xu, Sifei Liu, and Xiaolong Wang, "Semi-supervised 3d hand-object poses estimation with interactions in time," in *CVPR*, 2021.

[36] Lixin Yang, Kailin Li, Xinyu Zhan, Jun Lv, Wenqiang Xu, Jiefeng Li, and Cewu Lu, "Artiboost: Boosting articulated 3d hand-object pose estimation via online exploration and synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[37] Xingyu Chen, Yufeng Liu, Dong Yajiao, Xiong Zhang, Chongyang Ma, Yanmin Xiong, Yuan Zhang, and Xiaoyan Guo, "Mobrecon: Mobile-friendly hand mesh reconstruction from monocular image," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, 2017.

[39] Alejandro Newell, Kaiyu Yang, and Jia Deng, "Stacked hourglass networks for human pose estimation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

[40] Adrian Spurr, Umar Iqbal, Pavlo Molchanov, Otmar Hilliges, and Jan Kautz, "Weakly supervised 3d hand pose estimation via biomechanical constraints," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2020.